Scene semantics from long-term observation of people.

Jacob Menashe

October 5, 2012

Introduction

- Function over form.

Introduction

- Function over form.
- Form can be unique, function can be descriptive.

Introduction

- Function over form.
- Form can be unique, function can be descriptive.
- Learn semantics through observation.

Why semantics?

Semantics are great for...

Why semantics?

Semantics are great for...

- Abnormal event detection.

Why semantics?

Semantics are great for...

- Abnormal event detection.
- Event prediction.

Why semantics?

Semantics are great for...

- Abnormal event detection.
- Event prediction.
- Security and Surveillance

Why semantics?

Semantics are great for...

- Abnormal event detection.
- Event prediction.
- Security and Surveillance
- Achieving semantic objectives.

Why semantics?

Semantics are great for...

- Abnormal event detection.
- Event prediction.
- Security and Surveillance
- Achieving semantic objectives.
  - Robotics - performing tasks.

Why semantics?

Semantics are great for...

- Abnormal event detection.
- Event prediction.
- Security and Surveillance
- Achieving semantic objectives.
  - Robotics - performing tasks.
  - Database search - offering suggestions.

Related Work

- Semantic labeling on outdoor scenes: Kohli and Torr [2008], Shotton et al. [2006].

Related Work

- Semantic labeling on outdoor scenes: Kohli and Torr [2008], Shotton et al. [2006].
- Action recognition on still images: Gupta et al. [2009], Delaitre et al. [2011].

Related Work

- Semantic labeling on outdoor scenes: Kohli and Torr [2008], Shotton et al. [2006].
- Action recognition on still images: Gupta et al. [2009], Delaitre et al. [2011].
- Object localization on still images: Gupta et al. [2009], Desai et al. [2010], Stark et al. [2008].

Related Work

- Semantic labeling on outdoor scenes: Kohli and Torr [2008], Shotton et al. [2006].
- Action recognition on still images: Gupta et al. [2009], Delaitre et al. [2011].
- Object localization on still images: Gupta et al. [2009], Desai et al. [2010], Stark et al. [2008].
- Pose estimation on still images: Yao and Fei-fei [2010], Yao et al. [2011].

Related Work

- Semantic labeling on outdoor scenes: Kohli and Torr [2008], Shotton et al. [2006].
- Action recognition on still images: Gupta et al. [2009], Delaitre et al. [2011].
- Object localization on still images: Gupta et al. [2009], Desai et al. [2010], Stark et al. [2008].
- Pose estimation on still images: Yao and Fei-fei [2010], Yao et al. [2011].
- Coarse functional descriptions for surveillance: Peursum et al. [2005], Turek et al. [2010], Wang et al. [2006].

Related Work

- Semantic labeling on outdoor scenes: Kohli and Torr [2008], Shotton et al. [2006].
- Action recognition on still images: Gupta et al. [2009], Delaitre et al. [2011].
- Object localization on still images: Gupta et al. [2009], Desai et al. [2010], Stark et al. [2008].
- Pose estimation on still images: Yao and Fei-fei [2010], Yao et al. [2011].
- Coarse functional descriptions for surveillance: Peursum et al. [2005], Turek et al. [2010], Wang et al. [2006].
- Functions or affordances from 3D Reconstructions: Grabner et al. [2011], Gupta et al. [2011], Gibson [1979].

Overview

Time-lapse video

*

Image taken from Delaitre et al. [2012]

Overview

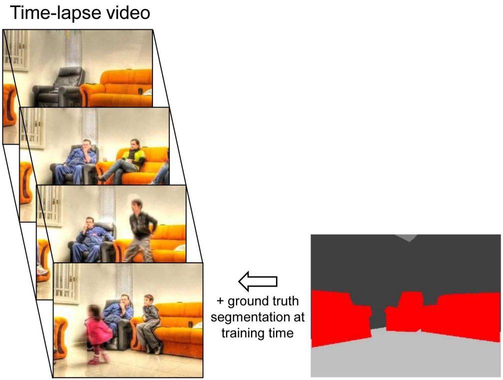


Image taken from Delaitre et al. [2012]

Overview

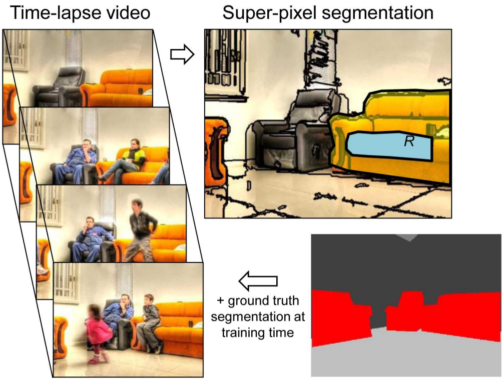


Image taken from Delaitre et al. [2012]

Overview

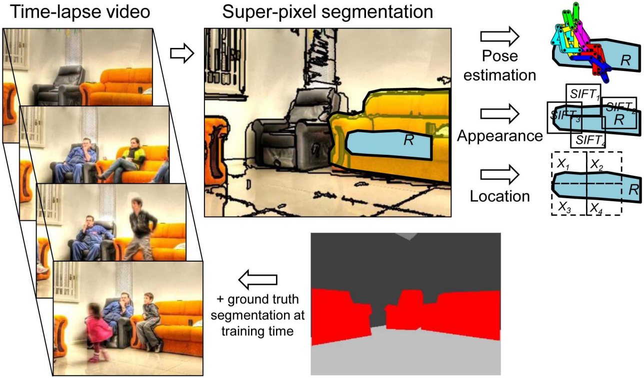
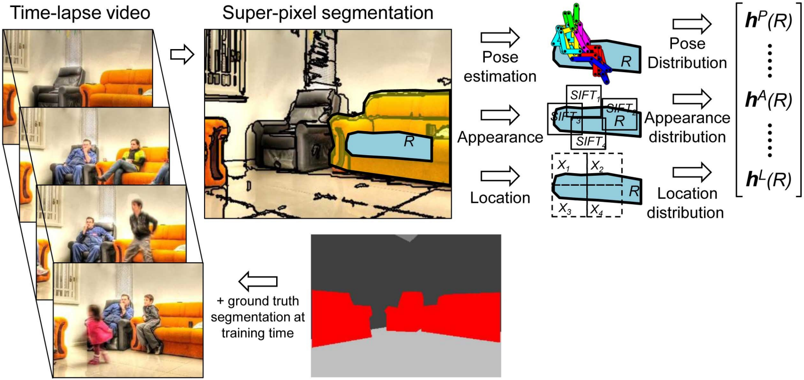


Image taken from Delaitre et al. [2012]

Overview



*

Image taken from Delaitre et al. [2012]

Overview

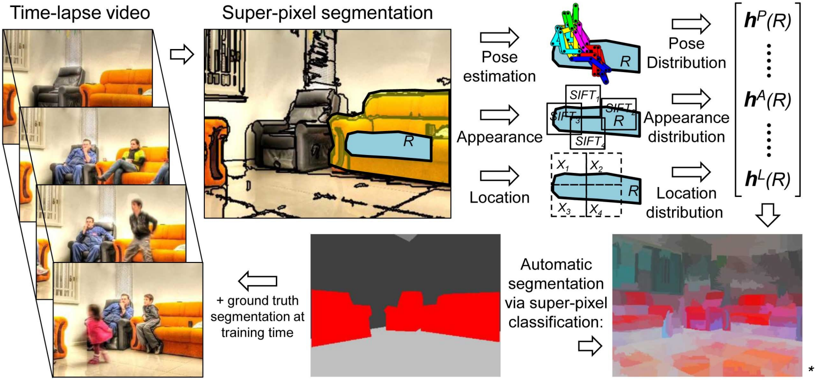
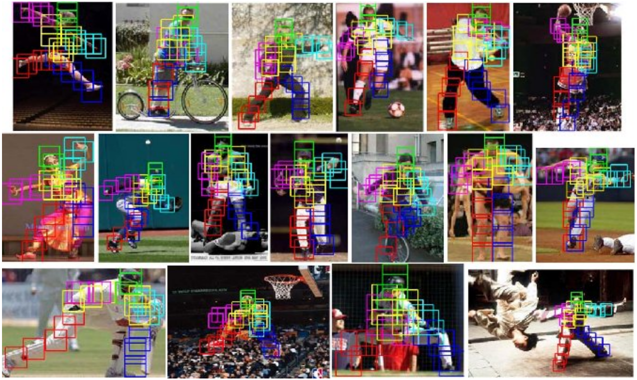


Image taken from Delaitre et al. [2012]

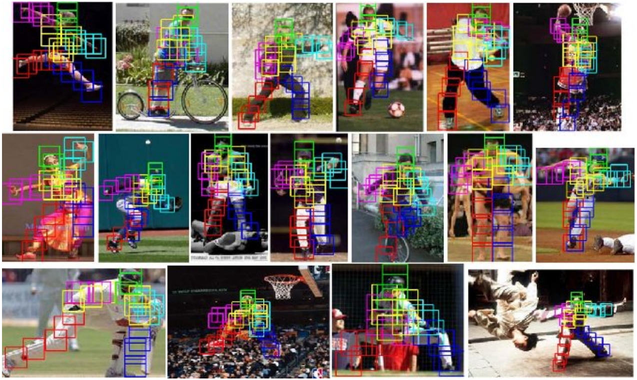
Pose Detection

- Pose detection begins with the person detector from Yang and Ramanan [2011].



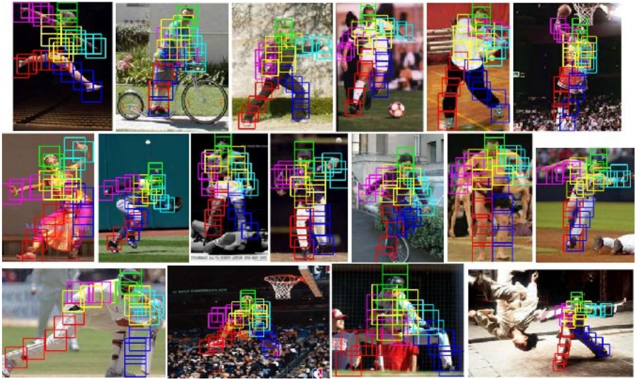
Pose Detection

- Pose detection begins with the person detector from Yang and Ramanan [2011].
- 3 Models, 3 detectors, merged into 1.



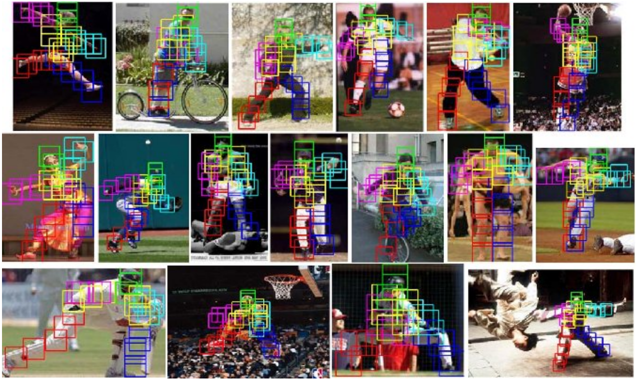
Pose Detection

- Pose detection begins with the person detector from Yang and Ramanan [2011].
- 3 Models, 3 detectors, merged into 1.
- Standing



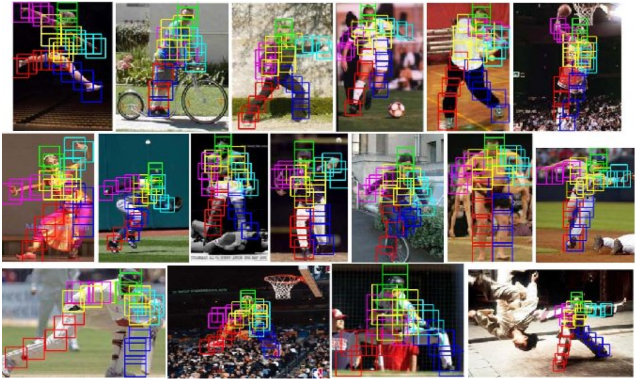
Pose Detection

- Pose detection begins with the person detector from Yang and Ramanan [2011].
- 3 Models, 3 detectors, merged into 1.
- Standing
- Sitting



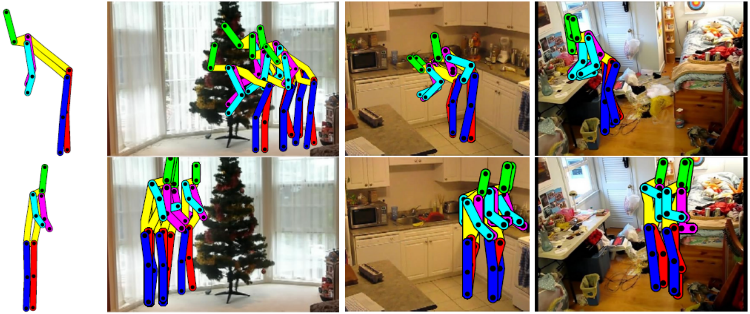
Pose Detection

- Pose detection begins with the person detector from Yang and Ramanan [2011].
- 3 Models, 3 detectors, merged into 1.
- Standing
- Sitting
- Reaching

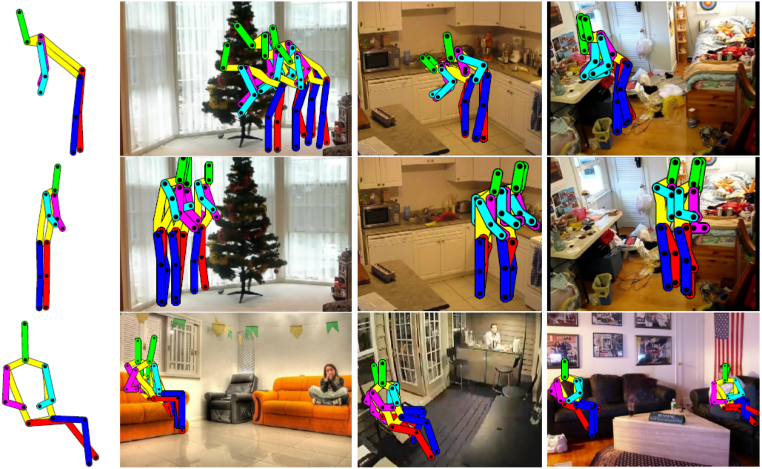


Pose Model

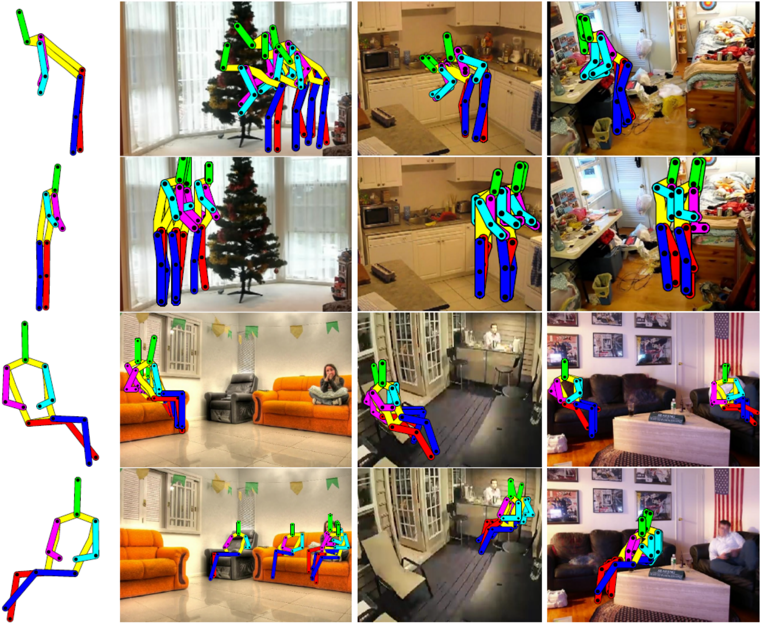
Pose Model



Pose Model



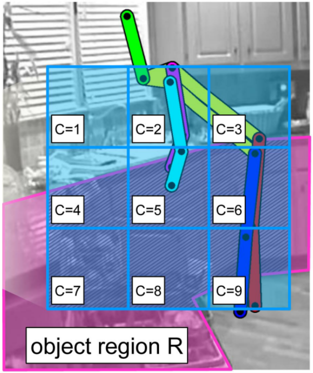
Pose Model



Relative Object Location

- Joint-region overlaps are determined.

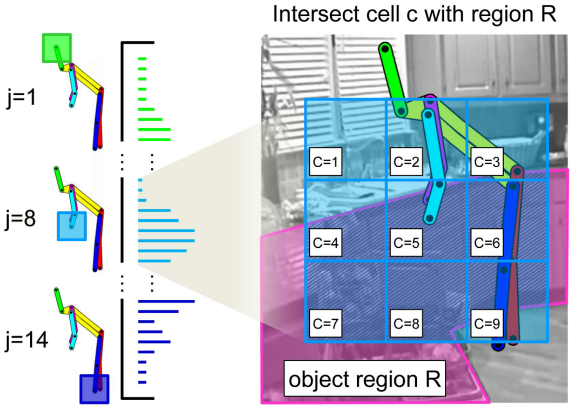
Intersect cell c with region R



* *Image taken from Delaitre et al. [2012]*

Relative Object Location

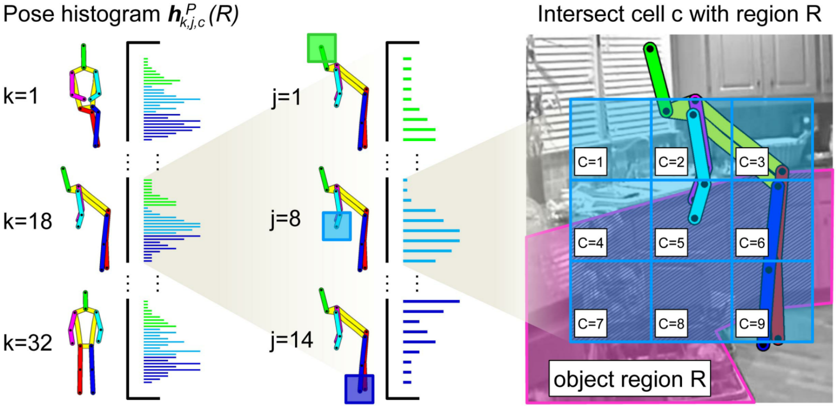
- Joint-region overlaps are determined.
- Overlaps are aggregated.



** Image taken from Delaitre et al. [2012]*

Relative Object Location

- Joint-region overlaps are determined.
- Overlaps are aggregated.
- Histograms are weighted by pose likelihood.



* *Image taken from Delaitre et al. [2012]*

Pose Histogram

$$h^{P}_{k,j,c}(R) = \sum_{d \in \mathcal{D}} \frac{\mathcal{I}(B_{j,c}, R)}{1 + \exp\left(-3 s_d\right)} q^{d}_{k}$$

Pose Histogram

$$h^P_{k,j,c}(R) = \sum_{d \in \mathcal{D}} \frac{\mathcal{I}(B_{j,c}, R)}{1 + \exp\left(-3 s_d\right)} q^d_k$$

- $\mathcal{D}$ is the detections.

Pose Histogram

$$h^{P}_{k,j,c}(R) = \sum_{d \in \mathcal{D}} \frac{\mathcal{I}(B_{j,c}, R)}{1 + \exp\left(-3 s_d\right)} q^{d}_{k}$$

- $\mathcal{D}$ is the detections.
- s_d is the score of the detection.

Pose Histogram

$$h^{P}_{k,j,c}(R) = \sum_{d\in\mathcal{D}} \frac{\mathcal{I}(B_{j,c}, R)}{1 + \exp\left(-3s_d\right)} q^d_k$$

- $\mathcal{D}$ is the detections.
- s_d is the score of the detection.
- q^d_k is the pose assignment coefficient for pose k.

Pose Histogram

$$h^{P}_{k,j,c}(R) = \sum_{d \in \mathcal{D}} \frac{\mathcal{I}(B_{j,c}, R)}{1 + \exp(-3 s_d)} q^d_k$$

- $\mathcal{D}$ is the detections.
- s_d is the score of the detection.
- q^d_k is the pose assignment coefficient for pose k.
- $\mathcal{I}$ is the intersection.

Object Appearance Model

Object appearances are modeled with bag-of-words.

Appearance Histogram

$$\boldsymbol{h}^A(R) = \sum_{k=1}^{S} \sum_{f \in \mathcal{F}_k} s_k^2 \mathcal{I}(B^f, R)\boldsymbol{q}^f$$

- $\mathcal{F}_k$ is the SIFT features.

Appearance Histogram

$$\boldsymbol{h}^A(R) = \sum_{k=1}^{S} \sum_{f \in \mathcal{F}_k} s_k^2 \mathcal{I}(B^f, R) \boldsymbol{q}^f$$

- $\mathcal{F}_k$ is the SIFT features.
- s_k is the window size.

Appearance Histogram

$$\boldsymbol{h}^A(R) = \sum_{k=1}^{S} \sum_{f \in \mathcal{F}_k} s_k^2 \mathcal{I}(B^f, R) \boldsymbol{q}^f$$

- $\mathcal{F}_k$ is the SIFT features.
- s_k is the window size.
- $\mathcal{I}(B^f, R)$ is region-box intersection.

Appearance Histogram

$$\boldsymbol{h}^A(R) = \sum_{k=1}^{S} \sum_{f \in \mathcal{F}_k} s_k^2 \mathcal{I}(B^f, R) \boldsymbol{q}^f$$

- $\mathcal{F}_k$ is the SIFT features.
- s_k is the window size.
- $\mathcal{I}(B^f, R)$ is region-box intersection.
- $\boldsymbol{q}^f$ is the soft bag-of-words assignment.

Location Model

Finally, we model location data.

Location Model

Finally, we model location data.

- Discretize the video frame into cells.

Location Model

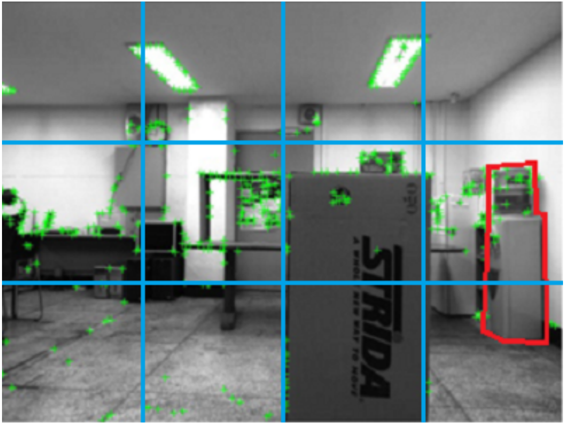
Finally, we model location data.

- Discretize the video frame into cells.
- $h_i^L(R)$ is the proportion of pixels in cell i falling into region R.

Location Model

Finally, we model location data.

- Discretize the video frame into cells.
- $h_i^L(R)$ is the proportion of pixels in cell i falling into region R.



Candidate Object Detection

- Video frames are over-segmented into super-pixels.

Candidate Object Detection

- Video frames are over-segmented into super-pixels.
- “Background” frame is defined.

Candidate Object Detection

- Video frames are over-segmented into super-pixels.
- “Background” frame is defined.
- Repeat to reduce noise.

Learning Object Model

- An SVM is trained for each object class:

Learning Object Model

- An SVM is trained for each object class:
  - Interactive - Bed, Sofa/Armchair, Coffee Table, Chair, Table, Wardrobe/Cupboard, Christmas tree, Other

Learning Object Model

- An SVM is trained for each object class:
  - Interactive - Bed, Sofa/Armchair, Coffee Table, Chair, Table, Wardrobe/Cupboard, Christmas tree, Other
  - Background - Wall, Ceiling, Floor

Learning Object Model

- An SVM is trained for each object class:
  - Interactive - Bed, Sofa/Armchair, Coffee Table, Chair, Table, Wardrobe/Cupboard, Christmas tree, Other
  - Background - Wall, Ceiling, Floor
- Each classifier is binary.

Inferring Probable Pose

- Objective: Choose a likely pose for a given area.

Inferring Probable Pose

- Objective: Choose a likely pose for a given area.
- Choose a pose cluster to maximize:

$$\hat{k} = \arg\max_{k} \sum_{j=1}^{J} \sum_{c=1}^{9} \sum_{\text{pixels } i \in B^{k}_{j,c}} w_{y_i}(k, j, c)$$

Inferring Probable Pose

- Objective: Choose a likely pose for a given area.
- Choose a pose cluster to maximize:

$$\hat{k} = \arg\max_{k} \sum_{j=1}^{J} \sum_{c=1}^{9} \sum_{\text{pixels } i \in B_{j,c}^{k}} w_{y_i}(k, j, c)$$

- k is the pose

Inferring Probable Pose

- Objective: Choose a likely pose for a given area.
- Choose a pose cluster to maximize:

$$\hat{k} = \arg\max_{k} \sum_{j=1}^{J} \sum_{c=1}^{9} \sum_{\text{pixels } i \in B_{j,c}^{k}} w_{y_i}(k, j, c)$$

- k is the pose
- j is the joint

Inferring Probable Pose

- Objective: Choose a likely pose for a given area.
- Choose a pose cluster to maximize:

$$\hat{k} = \arg\max_{k} \sum_{j=1}^{J} \sum_{c=1}^{9} \sum_{\text{pixels } i \in B_{j,c}^{k}} w_{y_i}(k, j, c)$$

- k is the pose
- j is the joint
- c is the joint cell

Inferring Probable Pose

- Objective: Choose a likely pose for a given area.
- Choose a pose cluster to maximize:

$$\hat{k} = \arg\max_{k} \sum_{j=1}^{J} \sum_{c=1}^{9} \sum_{\text{pixels } i \in B_{j,c}^{k}} w_{y_i}(k, j, c)$$

- k is the pose
- j is the joint
- c is the joint cell
- $B_{j,c}^{k}$ is the bounding box

Inferring Probable Pose

- Objective: Choose a likely pose for a given area.
- Choose a pose cluster to maximize:

$$\hat{k} = \arg\max_{k} \sum_{j=1}^{J} \sum_{c=1}^{9} \sum_{\text{pixels } i \in B_{j,c}^{k}} w_{y_i}(k, j, c)$$

- k is the pose
- j is the joint
- c is the joint cell
- $B_{j,c}^{k}$ is the bounding box
- $w_{y_i}(k, j, c)$ is the learned SVM weights for k, j, c in $\tilde{\boldsymbol{h}}^{P}(R)$.

Annotated Video Datasets

- ~150 time-lapse videos of indoor environments

Annotated Video Datasets

- ~150 time-lapse videos of indoor environments
- Stationary cameras

Annotated Video Datasets

- ~150 time-lapse videos of indoor environments
- Stationary cameras
- Manual annotation of single frames

Annotated Video Datasets

- ~150 time-lapse videos of indoor environments
- Stationary cameras
- Manual annotation of single frames
- `http://www.youtube.com/watch?v=17HXRdVzsrM`

Semantic Labeling

Labelings are evaluated with AP score.

	DPM [1]	Alternate [2]	(A + L)	(P)	(A + P)	(A + L + P)
Wall	-	75	76	76	**82**	81
Ceiling	-	47	53	52	**69**	**69**
Floor	-	59	64	65	**76**	**76**
Bed	**31**	12	14	21	27	26
Sofa/Armchar	26	26	34	32	**44**	43
Coffee Table	11	11	11	12	**17**	**17**
Chair	9.5	6.3	8.3	5.8	11	**12**
Table	15	18	17	16	**22**	**22**
Wardrobe/Cupboard	27	27	28	22	**36**	**36**
Christmas Tree	50	55	72	20	76	**77**
Other Object	12	11	7.9	13	**16**	**16**
Average	23	31	35	30	**43**	**43**

[1] Felzenszwalb et al. [2010]

[2] Hedau et al. [2009]

Semantic Labeling

Labelings are evaluated with AP score.

- Measured against two competing methods.

	DPM [1]	Alternate [2]	(A + L)	(P)	(A + P)	(A + L + P)
Wall	-	75	76	76	**82**	81
Ceiling	-	47	53	52	**69**	**69**
Floor	-	59	64	65	**76**	**76**
Bed	**31**	12	14	21	27	26
Sofa/Armchar	26	26	34	32	**44**	43
Coffee Table	11	11	11	12	**17**	**17**
Chair	9.5	6.3	8.3	5.8	11	**12**
Table	15	18	17	16	**22**	**22**
Wardrobe/Cupboard	27	27	28	22	**36**	**36**
Christmas Tree	50	55	72	20	76	**77**
Other Object	12	11	7.9	13	**16**	**16**
Average	23	31	35	30	**43**	**43**

[1] Felzenszwalb et al. [2010]

[2] Hedau et al. [2009]

Semantic Labeling

Labelings are evaluated with AP score.

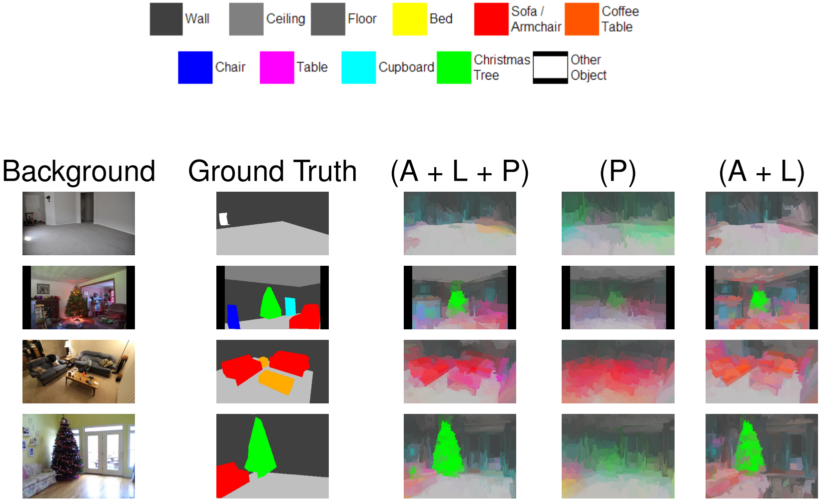
- Measured against two competing methods.
- (A+P), (A + L + P) outperform in all cases except for bed detection.

	DPM [1]	Alternate [2]	(A + L)	(P)	(A + P)	(A + L + P)
Wall	-	75	76	76	**82**	81
Ceiling	-	47	53	52	**69**	**69**
Floor	-	59	64	65	**76**	**76**
Bed	**31**	12	14	21	27	26
Sofa/Armchar	26	26	34	32	**44**	43
Coffee Table	11	11	11	12	**17**	**17**
Chair	9.5	6.3	8.3	5.8	11	**12**
Table	15	18	17	16	**22**	**22**
Wardrobe/Cupboard	27	27	28	22	**36**	**36**
Christmas Tree	50	55	72	20	76	**77**
Other Object	12	11	7.9	13	**16**	**16**
Average	23	31	35	30	**43**	**43**

[1] Felzenszwalb et al. [2010]

[2] Hedau et al. [2009]

Semantic Labeling Output



Functional Surface Estimation

- Measured with AP on functional labels

Functional Surface Estimation

- Measured with AP on functional labels
  - Walkable: 76%

Functional Surface Estimation

- Measured with AP on functional labels
  - Walkable: 76%
  - Sittable: 25%

Functional Surface Estimation

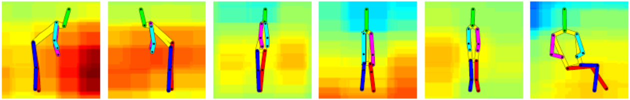
- Measured with AP on functional labels
  - Walkable: 76%
  - Sittable: 25%
  - Reachable: 44%

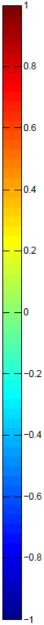
Functional Surface Estimation

- Measured with AP on functional labels
  - Walkable: 76%
  - Sittable: 25%
  - Reachable: 44%
- Average gain of 13% above baseline competitor: Fouhey et al. [2012]

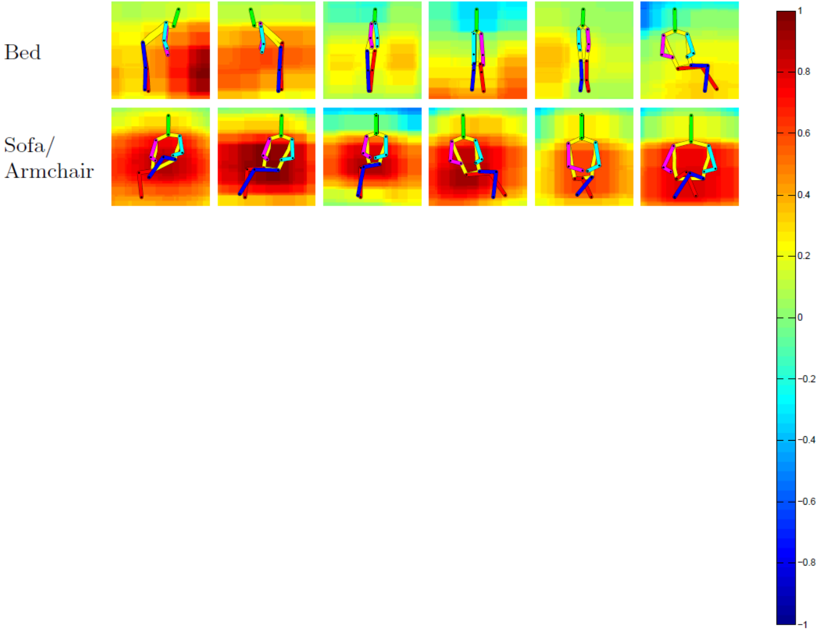
Pose-Region Relationships

Bed

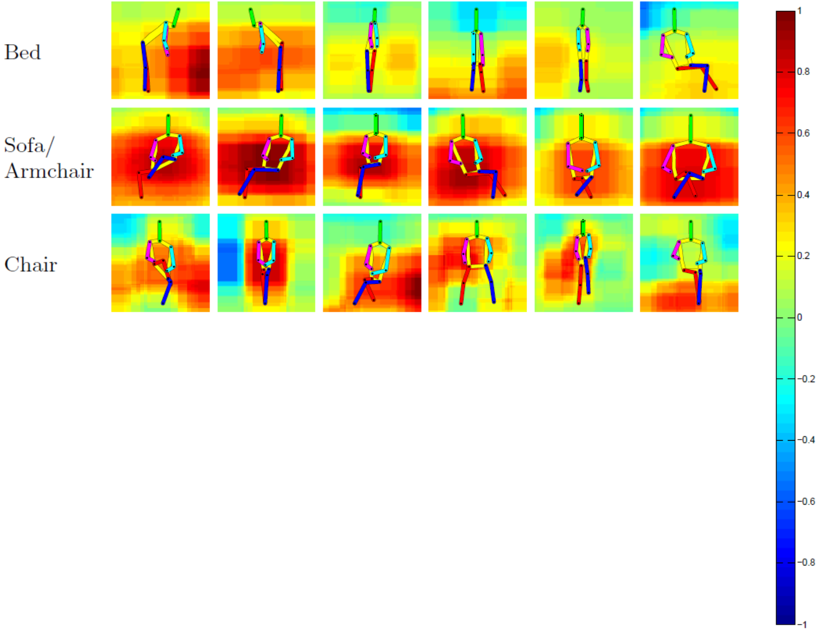




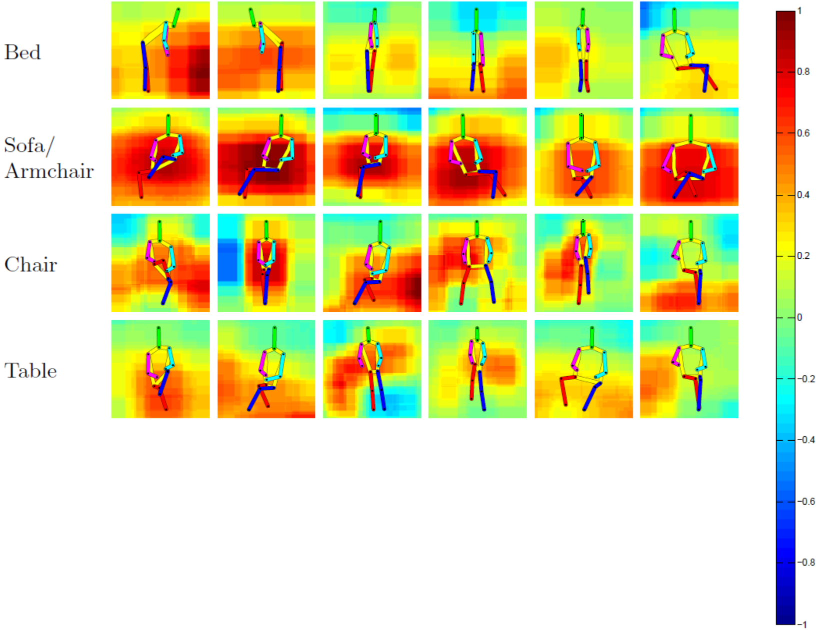
Pose-Region Relationships



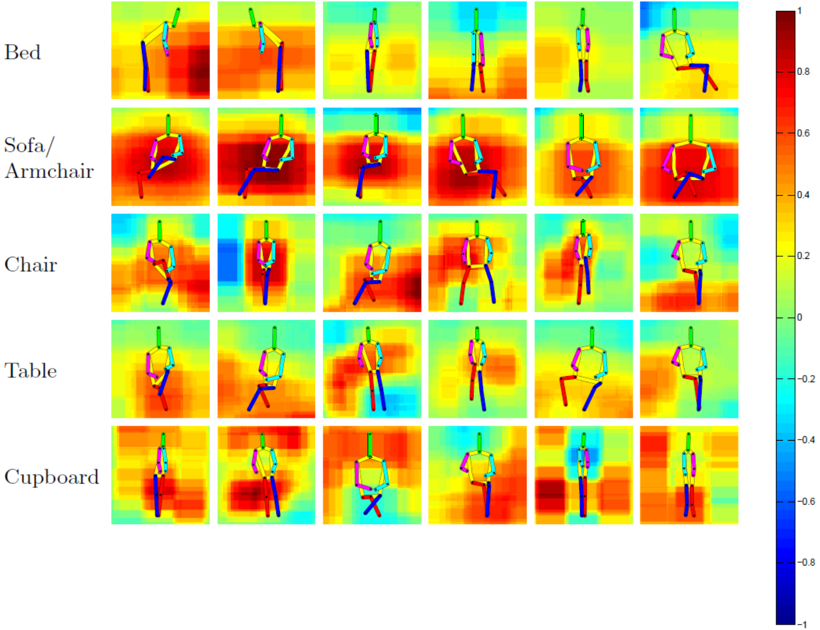
Pose-Region Relationships



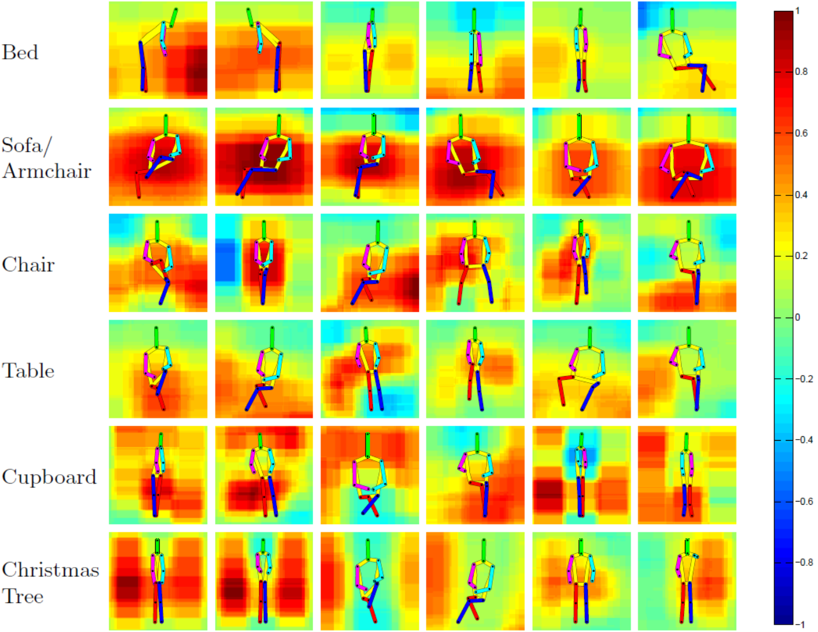
Pose-Region Relationships



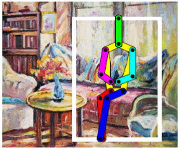
Pose-Region Relationships



Pose-Region Relationships

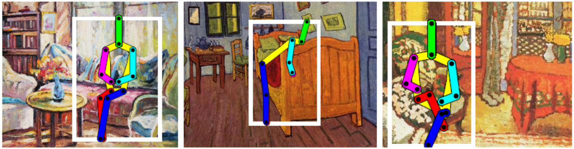


Pose Prediction

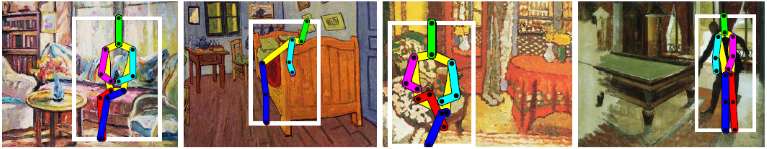


Pose Prediction

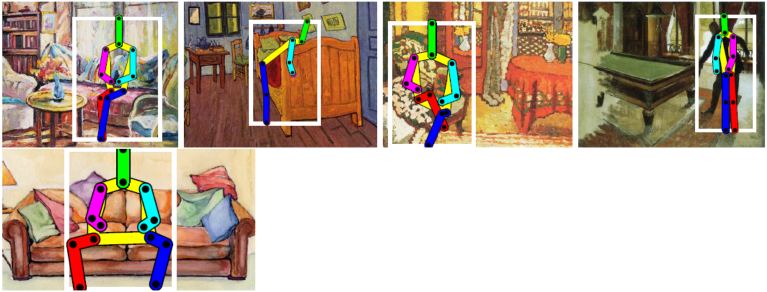
Pose Prediction



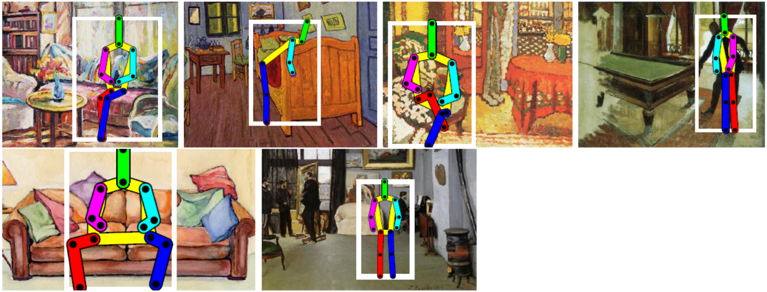
Pose Prediction



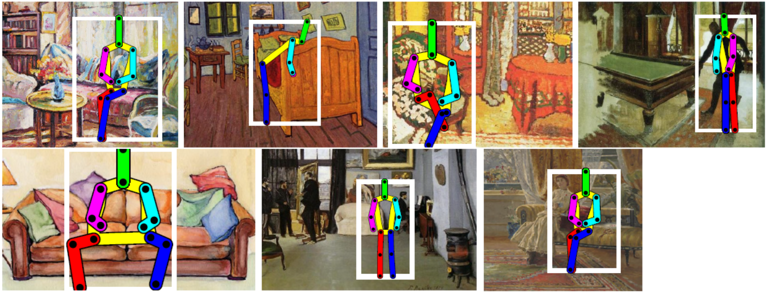
Pose Prediction



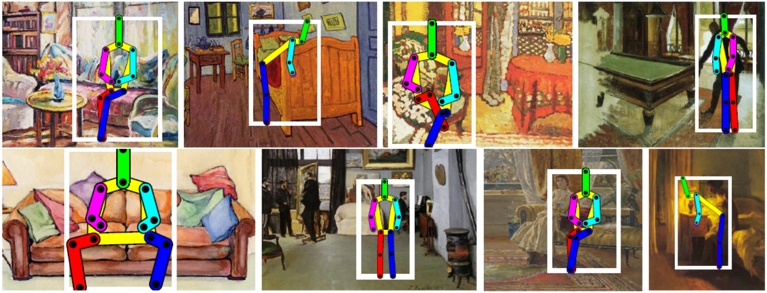
Pose Prediction



Pose Prediction



Pose Prediction



Extensions

- Using semantics as probabilistic information

Extensions

- Using semantics as probabilistic information
- Learning new objects from observation

Criticisms

Criticisms

- Lots of frames to learn a scene

Criticisms

- Lots of frames to learn a scene
- Weak precision rates

Criticisms

- Lots of frames to learn a scene
- Weak precision rates
- Manual annotations required

Conclusion

- Use observations to learn semantics.

Conclusion

- Use observations to learn semantics.
- Classify by semantic value.

Conclusion

- Use observations to learn semantics.
- Classify by semantic value.
- General enhancement to common detection systems.

References I

V. Delaitre, J. Sivic, and I. Laptev. Learning person-object interactions for action recognition in still images. In *Advances in Neural Information Processing Systems*, 2011.

V. Delaitre, D. Fouhey, I. Laptev, J. Sivic, A. Gupta, and A. Efros. Scene semantics from long-term observation of people. In *Proc. 12th European Conference on Computer Vision*, 2012.

Chaitanya Desai, Deva Ramanan, and Charless Fowlkes. C.: Discriminative models for static human-object interactions. In *In: Workshop on Structured Models in Computer Vision*, 2010.

Pedro F. Felzenszwalb, Ross B. Girshick, David A. McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(9):1627–1645, 2010.

References II

David F. Fouhey, Vincent Delaitre, Abhinav Gupta, Alexei A. Efros, Ivan Laptev, and Josef Sivic. People watching: Human actions as a cue for single-view geometry. In *Proc. 12th European Conference on Computer Vision*, 2012.

James Jerome Gibson. The ecological approach to visual perception. 1979. Houghton Mifflin.

Helmut Grabner, Juergen Gall, and Luc J. Van Gool. What makes a chair a chair? In *CVPR*, pages 1529–1536. IEEE, 2011. URL `http://dblp.uni-trier.de/db/conf/cvpr/cvpr2011.html#GrabnerGG11`.

A. Gupta, S. Satkin, A. A. Efros, and M. Hebert. From 3d scene geometry to human workspace. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '11, pages 1961–1968, Washington, DC, USA, 2011. IEEE Computer Society. ISBN 978-1-4577-0394-2. doi: 10.1109/CVPR.2011.5995448. URL `http://dx.doi.org/10.1109/CVPR.2011.5995448`.

References III

Abhinav Gupta, Aniruddha Kembhavi, and Larry S. Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31:1775–1789, 2009. ISSN 0162-8828. doi: http://doi.ieeecomputersociety.org/10.1109/TPAMI.2009.83.

Varsha Hedau, Derek Hoiem, and David Forsyth. Recovering the spatial layout of cluttered rooms, 2009.

Pushmeet Kohli and Philip H. S. Torr. Robust higher order potentials for enforcing label consistency. In *In CVPR*, 2008.

Patrick Peursum, Geoff West, and Svetha Venkatesh. Combining image regions and human activity for indirect object recognition in indoor wide-angle views. In *Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1 - Volume 01*, ICCV '05, pages 82–89, Washington, DC, USA, 2005. IEEE Computer

References IV

Society. ISBN 0-7695-2334-X-01. doi: 10.1109/ICCV.2005.57. URL http://dx.doi.org/10.1109/ICCV.2005.57.

J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost: Joint appearance, shape and context modeling for multi-class object . . . In *IN ECCV*, pages 1–15, 2006.

Michael Stark, Philipp Lies, Michael Zillich, Jeremy L. Wyatt, and Bernt Schiele. Functional object class detection based on learned affordance cues. In *6th International Conference on Computer Vision Systems (ICVS)*, Santorini, Greece, 2008 2008. URL http://www.mis.informatik.tu-darmstadt.de/People/stark/stark08icvs.pdf. Oral presentation.

References V

Matthew Turek, Anthony Hoogs, and Roderic Collins. Unsupervised learning of functional categories in video scenes. In Kostas Daniilidis, Petros Maragos, and Nikos Paragios, editors, *Computer Vision ECCV 2010*, volume 6312 of *Lecture Notes in Computer Science*, pages 664–677. Springer Berlin / Heidelberg, 2010. ISBN 978-3-642-15551-2. 10.1007/978-3-642-15552-9_48.

Xiaogang Wang, Kinh Tieu, and Eric Grimson. Learning semantic scene models by trajectory analysis. In *In ECCV (3) (2006*, pages 110–123, 2006.

Yi Yang and Deva Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, pages 1385–1392. IEEE, 2011.

Bangpeng Yao and Li Fei-fei. L.: Modeling mutual context of object and human pose in human-object interaction activities, 2010.

References VI

Bangpeng Yao, Aditya Khosla, and Li Fei-fei. Classifying actions and measuring action similarity by modeling the mutual context of objects and human poses. In *In Proc. ICML*, 2011.